

METHOD AND APPARATUS FOR FORECASTING AGENT ARRIVAL

FIELD OF THE INVENTION

5 The present invention is related to forecasting agent arrival. In particular, the present invention is related to forecasting the future availability of an agent for a new work assignment.

BACKGROUND OF THE INVENTION

 A predictive dialer is a system that generates outbound calls from a call center. Outbound calls that reach a live called party (i.e., a human being answers the call) are connected to call center agents for handling, i.e., servicing. Such answered calls for which no agent is immediately available to service them are referred to as nuisance calls. Nuisance calls are possible because the predictive dialer initiates calls without a priori assignment and connection of agents to the generated calls. Furthermore, because generated calls can go unanswered or may be answered by a machine as opposed to a live called party, the predictive dialer typically initiates more outbound calls than there are agents available to service them. With automatic call classification of whether or not, and by what, an outbound call has been answered, the call center determines the need for an agent only when the call is answered. The call center then attempts to match and connect the call answered by a live party to an individual agent for immediate servicing.

 A call center has the conflicting objectives of minimizing agent idle time while minimizing the numbers of nuisance calls. But two characteristics of telephone calling are inherently problematic for agent utilization. The first is the long ring time required to

reach a called party, and the second is a high incidence of call attempts that result in failure to connect to a live called party. Without predictive dialing, agents would spend considerable time waiting for ringing phones to be answered. When a ringing phone is not answered, the agent would have to wait again for results of the next call attempt.

- 5 Predictive dialing attempts to reduce agent wait-times. However, reducing agent wait-times is not without the risk of connecting to a called party without an agent being available to service the call.

10 In a predictive dialing environment, agents are presently unavailable because they are performing tasks that are designated to be un-interruptible. That is, an agent can do only one task at a time and cannot start a new task until the previous task is completed. Additionally, only one agent can usually be assigned to a particular task. The tasks typically involve live clients in a transaction, and background fulfillment tasks may be dynamically interspersed amongst them. Agents that are presently handling tasks are rendered unavailable. However, each unavailable agent can be expected to complete his
15 or her task with a particular probability over a future time interval.

Existing predictive dialers use various methods and algorithms to systematically initiate calls before agents are available to handle them. Whatever is the basic method that a dialer employs, it attempts to achieve a balance between agent utilization and nuisance-call rates. Such balancing typically employs heuristics realized in the predictive
20 dialer's dialing algorithm. For example, a dialer may consider an agent available after the agent has spent a certain amount of time working a call, and initiate new calls depending on the dialer's hit rate. For instance, if the hit rate is $1/3$, 3 calls will be initiated so that a live person can be expected to have been reached at about the time an

agent is predicted to be available. Accordingly, predictive dialers have been oriented towards keeping agents busy. However, dialers face many dynamic and unsystematic variations in characteristics of the called population, characteristics of the calling campaign, spontaneous assignment of agents to other work, and inconsistent behavior of agents. These variations tend to upset any real balance between agent utilization and nuisance-call rates. The existing predictive methods do not sufficiently compensate for the unsystematic uncertainty in their predictive models. Either there is excessive agent idle time (detected as an excess service level), or there is lack of promptness in responding to called parties who answer (detected as an insufficient service level). A service level is a percentage or a ratio of those transactions out of all transactions that satisfy some criterion of “goodness.”

One technique used by predictive dialers has been to consider the probability that an agent will finish a task within a selected amount of time based on the amount of time that the agent has spent on a call the agent is presently handling. However, such systems fail to consider correlations that may exist between talk time and wrap-up time. For example, such systems may result in an unrealistically high probability that an agent will become available within a selected period of time when the agent’s time in talk state becomes longer than normal. Other systems have simply considered the average time to advance in a wait queue, without consideration of the agents’ states. Because of the deficiencies in previous systems, the consistent achievement of higher service levels (*i.e.* the avoidance of nuisance calls), for example 99% of calls made to a live person are connected to an agent that is immediately available, has been impossible.

As a result of increasing concern about nuisance calls, regulations attempting to prevent or limit the occurrence of such calls have been implemented. These regulations can impose fines for calls that are completed to a live person but that are not also immediately connected to an agent. At the same time, it remains important to keep agents busy. Therefore, it has become increasingly important to improve the accuracy of predictive dialers to avoid instances of nuisance calls while at the same time avoiding having agents idle.

SUMMARY OF THE INVENTION

The present invention is directed to solving these and other problems and disadvantages of the prior art. According to embodiments of the present invention, the probability of completing a task that can be broken into a number of differentiated segments is determined by determining the probabilities with respect to such differentiated segments. In accordance with further embodiments of the present invention, the probability of completing a task determined with respect to the entire task is combined with the probability obtained by determining the probabilities with respect to the differentiated segments. The probability of completing a task may be determined with reference to a selected time horizon or forecast horizon.

In accordance with exemplary embodiments of the present invention, talk time statistics are used in combination with wrap-up time statistics to predict a time at which an agent will become available. Accordingly, by applying forecasts from a number of different models, greater granularity is available, allowing predictions to be made with improved accuracy. In accordance with further embodiments of the present invention, the predictions made using talk time statistics and wrap-up time statistics are combined

with total handle time statistics to make a prediction as to agent availability within a forecast horizon. The forecast time horizon generally depends on the context of the system. For example, the forecast time horizon may comprise a predicted time to a customer contact on the next call attempt.

5 In accordance with embodiments of the present invention, the probability of completing a talk state within a time horizon, and the probability of completing a wrap-up state within the time horizon, are multiplied by one another. Because the product of the multiplication results in a time horizon that is effectively twice that of the time horizon used to determine the probabilities, the product may be divided by two. In
10 accordance with embodiments of the present invention, the time horizon over which the probabilities are determined is selected to be equal to the mean time at which a live called party is expected to be available in response to a new call attempt.

 The statistics applied may be differentiated based on the type of task they present to an agent. For example, tasks may be differentiated based on the channel type, the
15 relationship between the agent and the other party associated with the task, a classification of the other party, the type of transaction, etc. Furthermore, embodiments of the present invention may be applied in connection with customer service applications, predictive dialing applications, or other applications in which a resource is engaged in a task that may be considered as a set of segments, and in which a forecast time horizon
20 within which a next resource will be required is made.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram depicting components of a system in accordance with embodiments of the present invention;

Fig. 2 is a flow chart depicting aspects of the operation of a system in accordance with embodiments of the present invention; and

Fig. 3 is a flow chart depicting aspects of the operation of another system in accordance with embodiments of the present invention.

DETAILED DESCRIPTION

With reference now to **Fig. 1**, a block diagram of a system 100 in accordance with embodiments of the present invention is illustrated. As shown in **Fig. 1**, the system 100 may include a call center including an automatic call distributor (ACD) 104 that interconnects agent positions 130-140 via calls with the outside world to which it is connected by communications trunks 102. In accordance with embodiments of the present invention, the communications trunks may comprise communications channels established over the public switched telephone network (PSTN), leased lines, the Internet, a private intranet, or any other communications networks, alone or in combination. ACD 104 is a stored program-controlled apparatus that operates under control of a processor 112 that obtains and stores data in, and executes stored programs out of, memory 110 or any other computer-readable medium. Programs in memory 110 include a call classifier, which analyzes the signals on a call connection to distinguish control signals from other signals (e.g., voice, music) and to determine their meaning. Data in memory 110 may include historical and operational data of the ACD 104 and of agents 156, which are stored in a call management system (CMS) database 120. Processor 112 controls the

operation of a switching fabric 116 and of a predictive dialer 114 that generates outgoing calls on the trunks 102 through the switching fabric 116. Each agent position 130-140 includes a terminal 152, such as a personal computer, and a voice communications device, such as a telephone or a headset 154, for use by an agent 156.

5 For purposes of the discussion herein, a call, whether incoming or outgoing, constitutes a task to be served, and an agent position 130-140 that is presently staffed by an agent 156 constitutes a resource for serving tasks. In addition, although certain examples provided herein are in the context of a call center for handling voice calls, the present invention is not so limited. For example, embodiments of the present invention
10 may be applied in connection with systems for distributing tasks, including real time tasks, to resources in which the duties of agents or other resources can be differentiated or segmented.

According to an illustrative embodiment of the invention, memory 110 of the ACD 104 includes a customer contact forecast function 122. The contact forecast
15 function 122 could alternatively be included and executed either in dialer 114 or in an adjunct processor (not shown) connected to the ACD 104. While implemented in software in the illustrative embodiment of Fig. 1, the customer contact forecast function 122 could alternatively be implemented in hardware or firmware.

The customer contact forecast function 122, together with the other functions of
20 the ACD 104, including the predictive dialer 114, operates to provide a forecast in the form of a probability of a client contact or a prediction of the time until a next client contact. For instance, the customer contact forecast function 122 determines the amount of time it takes to get a live disposition on an incremental outbound call. Accordingly,

the customer contact forecast function provides a time horizon in which a prediction as to agent availability is to be made, as described in greater detail elsewhere herein.

In addition, the memory 110 of the ACD 104 may include an agent arrival prediction function 124. The agent arrival prediction function 124 could alternatively be included and executed either in dialer 114 or in an adjunct processor (not shown) connected to the ACD 104. While implemented in software in the illustrative embodiment of **Fig. 1**, the agent arrival prediction function 124 could alternatively be implemented in hardware or firmware.

In general, the agent arrival prediction function 124, together with certain of the other functions of the ACD 104, including the customer contact forecast function 122, operates to determine the probability that an agent 156 will be available within a forecast horizon. With reference now to **Fig. 2**, the operation of a system 100 incorporating an agent arrival prediction function 124 in accordance with an exemplary embodiment of the present invention is illustrated. Initially, at step 200, a forecast horizon is selected. In general, the forecast horizon may comprise a prediction as to the time until a next live disposition is expected in connection with the initiation of a call by a predictive dialer 114. In addition or alternatively, the forecast horizon may comprise a prediction as to the time until a next task requiring service from an agent 156, such as a request for service received as an inbound call from a customer, is received or generated.

At step 204, statistics regarding a task presently occupying an agent 156 associated with the system 100 that could possibly be eligible to service the customer contact are retrieved. Such statistics may be stored in the CMS database 120, or in another database associated with the ACD 104. According to embodiments of the present

invention, for each agent 156, the task occupying the agent is considered as two or more segments. The statistics generally comprise a probability distribution related to the availability of an agent engaged in a task segment with respect to time. For instance, the probability that an agent will have completed a task segment after a certain amount of time has elapsed may be expressed as a histogram or other probability distribution function. In addition, the statistics accessed may be composite statistics developed for agents generally, or particular statistics developed for the individual agent under consideration.

At step 208, the probability that the agent under consideration will complete the task in which the agent is engaged within the forecast horizon is determined by combining the probabilities for each of the two or more segments. As will be described in greater detail elsewhere herein, where the two or more segments are sequential, the actual time that the agent has spent within the first segment and the probability associated with a time of zero spent in the additional segment or segments are combined to obtain the probability for that agent. As will also be described in greater detail elsewhere herein, where the segments overlap one another, the actual time spent in each of the segments may be used to obtain a probability for each segment, and the probabilities thus obtained may then be combined to provide a probability as to whether the agent under consideration will become available within the forecast horizon.

A determination is then made as to whether additional eligible agents 156 are associated with the system 100 that need to have a probability determination made (step 212). If eligible agents 156 remain, the process returns to step 204. If all eligible agents 156 have been considered, the probabilities for each of those agents are combined to

obtain an overall expected value of the number of agents that will be available within the forecast horizon (step 216). Additionally, a variance of the number of agents that will be available may be obtained. Depending on the expected value and the variance in agent arrivals, the operation of the predictive dialer 114 may be modified, for example to ensure that the number of outgoing calls launched from the ACD 104 is not expected to result in more live dispositions than available agents 156.

With reference now to **Figs. 3A** and **3B**, the operation of a system 100 in accordance with a further exemplary embodiment is illustrated. Initially, and with particular reference to **Fig. 3A**, at step 300, a forecast time horizon is selected. For example, the selection of a forecast time horizon may comprise receiving from a predictive dialer 114 a forecast time until a live disposition on a call.

At step 304, talk time (i.e. a first task segment) statistics and wrap-up time (i.e. a second task segment) statistics are obtained for an agent 156, for the forecast horizon or interval. Using the time that the agent has actually been in talk state and the talk time statistics, a probability (or first probability) that the agent will complete the talk segment of a customer contact or service procedure is determined (step 308). As can be appreciated by one of skill in the art, this results in a value between one and zero for an arrival of the agent within the forecast horizon. In addition, a probability (or second probability) with respect to the completion of wrap-up within the forecast horizon is determined using zero for the time that the agent has been in wrap-up state (step 312). As can be appreciated by one of skill in the art, this gives a value between one and zero for an arrival of the agent within the forecast horizon. Alternatively, a probability that the

agent will arrive from wrap-up within the forecast horizon may be determined from a classical regression analysis of wrap-up time from talk time.

The probabilities for arrival of the agent from talk state (the first probability obtained in step 308) and wrap-up state (the second probability obtained in step 312) are then combined (step 316). In accordance with embodiments of the present invention, combining these probabilities comprises computing the product of the probabilities from steps 312 and 316. Alternatively, other combinations of the probability that the agent will arrive from talk state and wrap-up state may be made, such as simple addition, a weighted sum, and combinations of combinations. Where the product of the probabilities is used, the result is divided by two, in order to compensate for the effective doubling of the length of the forecast interval that results from using the same forecast interval in steps 308 and 312, to obtain the probability of arrival (p_{alt}) based on talk and wrap-up only (step 320). The step of dividing the result obtained by combining the talk time and wrap-up time statistics assumes that the distribution of occurrences would be uniform in a time interval twice the length of the forecast horizon. Assuming an exponential distribution, the factor would be $\frac{1}{1 + e^{-t_f/T}}$, where t_f is the length of the forecast time horizon and where T is the mean handle time. This factor approaches one-half as the forecast horizon becomes smaller compared to the mean handle time. Assumptions of other kinds of probability distributions are possible for the purpose of adjusting for the length of the forecast horizon to approximate the probability of arrival within the forecast horizon.

At step 324, total handle time statistics are obtained for an agent 156. From the total handle time statistics, a third probability or forecast, the probability of arrival from

the total handle time (p_w), with respect to the availability of the agent under consideration is determined (step 328). The ratio of the variance in the total handle time statistics to the sum of the variances in the talk time statistics, wrap-up time statistics, and total handle time statistics is then computed (step 332). In particular, the ratio is given as

$$\frac{V_{handle}}{V_{handle} + V_{talk} + V_{wu}}.$$

With reference now to **Fig. 3B**, at step 336, the a priori probability (F_{talk}) of completion of the talk state before the present time in talk state is obtained, for example from the talk time statistics obtained in step 304. The product of the ratio obtained at step 332 and the probability from step 336 is then computed as $w_{talk} = F_{talk} \frac{V_{handle}}{V_{handle} + V_{talk} + V_{wu}}$,

and used as a weight applied to the probability of arrival (p_{alt}) obtained in step 320, to obtain a first weighted forecast, given by $w_{talk} * p_{alt}$ (step 340). Alternatively, a weight based on correlations and coefficients of determination may be applied.

At step 344, a weight equal to one minus the weight applied in step 340 is applied to the forecast based on the total handle time statistics or probability of arrival from the total handle time (p_w) obtained at step 328 to obtain a second weighted forecast. This second weighted forecast is given by $(1 - w_{talk}) p_w$.

A composite forecast is then computed as the sum of the first and second weighted forecasts (step 348). In particular, the composite forecast or composite probability of arrival (p) is given by:

$$p = w_{talk} p_{alt} + (1 - w_{talk}) p_w.$$

The composite forecast may then be used to provide a probability that the agent under consideration will be available within the forecast horizon.

At step 352, a determination is made as to whether there are additional agents for which a forecast should be prepared. If there are additional agents, the process returns to step 304. If no agents remain, the composite probability of arrival for each of the agents is combined to obtain an overall forecast of the expected value of the number of agents that will be available within the forecast horizon (step 356). Additionally, the probabilities can be used to obtain a variance in the number of agents that will be available.

Embodiments of the present invention facilitate agent allocation and various dispatching operations, especially in an environment of differentiated work and differentiated agents where expeditious service is desired for at least some of the work items. For example, the integration of back-office work with front-office call work, including the integration of outbound call systems with workflow systems and automatic call distributors, can be facilitated. Furthermore, embodiments of the present invention allow a fusion of forecasts to be made for improved forecast accuracy, while avoiding the need to perform a convolution, which would be computationally expensive and complicated to implement.

A workflow system in accordance with embodiments of the present invention can also be used to ensure that an available agent is not assigned to a low priority work item when the agent should be dedicated to higher priority work, such as handling live outgoing or incoming telephone calls. In addition, embodiments of the present invention may be used to assign work to resources in contexts other than call centers featuring live agents. In general, embodiments of the present invention may be applied to distributing work to resources performing tasks that can be differentiated into a number of sub-tasks.

Furthermore, embodiments of the present invention allow high service levels and high agent or resource utilization. Accordingly, embodiments of the present invention may be used to provide improved customer service and to reduce the burden on supervision in call centers.

5 The foregoing discussion of the invention has been presented for purposes of illustration and description. Further, the description is not intended to limit the invention to the form disclosed herein. Consequently, variations and modifications commensurate with the above teachings, within the skill and knowledge of the relevant art, are within the scope of the present invention. The embodiments described hereinabove are further
10 intended to explain the best mode presently known of practicing the invention and to enable others skilled in the art to utilize the invention in such or in other embodiments with various modifications required by their particular application or use of the invention. It is intended that the appended claims be construed to include the alternative
 embodiments to the extent permitted by the prior art.